



## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : <b>H04Q 11/04, H04L 12/56</b>	A1	(11) International Publication Number: <b>WO 99/35879</b> (43) International Publication Date: 15 July 1999 (15.07.99)
(21) International Application Number: <b>PCT/US99/00607</b> (22) International Filing Date: 12 January 1999 (12.01.99)  (30) Priority Data: 09/005,738 12 January 1998 (12.01.98) US  (71) Applicant: CABLETRON SYSTEMS, INC. [US/US]; 35 Industrial Way, Rochester, NH 03867 (US).  (72) Inventors: CHARNY, Anna; 408 Dutton Road, Sudbury, MA 01776 (US). KRISHNA, Pattabhiraman; 21 Royal Crest Drive #9, Marlboro, MA 01752 (US). PATEL, Naimish; 32 Monteiro Way, N. Andover, MA 01845 (US). SIMCOE, Robert, J.; 11 Brookway Road, Westboro, MA 01581 (US).  (74) Agent: SORKIN, Paul, D.; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210 (US).	(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).  <b>Published</b> With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.	
<p><b>(54) Title:</b> METHOD FOR PROVIDING BANDWIDTH AND DELAY GUARANTEES IN A CROSSBAR SWITCH WITH SPEEDUP</p> <p><b>(57) Abstract</b></p> <p>An arbitration scheme for providing deterministic bandwidth and delay guarantees in an input-buffered crossbar switch with speedup <math>S</math> is presented. Within the framework of a crossbar architecture having a plurality of input channels and output channels, the arbitration scheme determines the sequence of fixed-size packet (or cell) transmission between the input channels and output channels satisfying the constraint that only one cell can leave an input channel and enter an output channel per phase in such a way that the arbitration delay is bounded for each cell awaiting transmission at the input channel. If the fixed-sized packets result from fragmentation of variable size packets, the scheduling and arbitration scheme determines deterministic delay guarantees to the initial variable size packets (re-assembled at the output channel) as well.</p>		

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon			PT	Portugal		
CN	China	KR	Republic of Korea	RO	Romania		
CU	Cuba	KZ	Kazakhstan	RU	Russian Federation		
CZ	Czech Republic	LC	Saint Lucia	SD	Sudan		
DE	Germany	LI	Liechtenstein	SE	Sweden		
DK	Denmark	LK	Sri Lanka	SG	Singapore		
EE	Estonia	LR	Liberia				

**METHOD FOR PROVIDING BANDWIDTH AND DELAY GUARANTEES IN A CROSSBAR SWITCH WITH SPEEDUP**

**FIELD OF THE INVENTION**

5       The present invention relates generally to variable and fixed size packet switches, and more particularly, to an apparatus and method for scheduling packet cell input through such packet switches.

**BACKGROUND OF THE INVENTION**

In the field of Integrated Services Networks, the importance of maintaining Quality of Service (QoS) guarantees for individual traffic streams (or flows) is generally recognized. Thus, such capability continues to be the subject of much research and development. Of particular interest for a system providing guaranteed flows are the guarantees associated with bandwidth and delay properties. These guarantees must be provided to all flows abiding to their service contract negotiated at connection setup, even in the presence of other potentially misbehaving flows. Many different methods have been developed to provide such guarantees in non-blocking switch architectures such as output buffered or shared memory switches. Several algorithms providing a wide range of delay guarantees for non-blocking architectures have been disclosed in the literature. See, for example, A. Parekh, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks", MIT, Ph.D dissertation, June 1994; J. Bennett and H. Zhang, "WF2Q - Worst-case Fair Weighted Fair Queuing", Proc. IEEE INFOCOM'96; D. Stiliadis and A. Varma, "Frame-Based Fair Queuing: A New Traffic Scheduling Algorithm for Packet Switch Networks", Proc. IEEE INFOCOM '96; L. Zhang, "A New Architecture for Packet Switched Network Protocols," Massachusetts Institute of Technology, Ph.D Dissertation, July 1989; A. Charny, "Hierarchical Relative Error Scheduler: An Efficient Traffic Shaper for Packet Switching Networks," Proc. NOSSDAV '97, May 1997, pp. 283-294; and others. Schedulers capable of providing bandwidth and delay guarantees in non-blocking architectures are commonly referred to as "QoS-capable schedulers".

Typically, output-buffered or shared memory non-blocking architectures require the existence of high-speed memory. For example, an output-buffered switch requires that the speed 30 of memory at each output must be equal to the total speed of all inputs. Unfortunately, the rate of the increase in memory speed available with current technology has not kept pace with the rapid growth in demand for providing large-scale integrated services networks. Because there is a growing demand for large switches with total input capacity of the order of tens and hundreds

of Gb/s, building an output buffered switch at this speed has become a daunting task given the present state of technology. Similar issues arise with shared memory switches as well.

As a result, many industrial and research architectures have adopted a more scalable approach, for example, crossbars. Details of such architectures may be had with reference to the  
5 following papers: T. Anderson, S. Owicki, J. Saxe, C. Thacker, "High Speed Switch Scheduling for Local Area Networks", Proc. Fifth Internat. Conf. on Architectural Support for Programming Languages and Operating Systems," Oct. 1992, pp. 98-110; and N. McKeown, M. Izzard, A. Mekkittikul, W. Ellersick and M. Horowitz, "The Tiny Tera: A Packet Switch Core." Even given the work already done, providing bandwidth and delays in an input-queued crossbar switch  
10 remains a significant challenge.

A paper by N. McKeown, V. Anatharam and J. Warland, entitled "Achieving 100% Throughput in an Input-Queued Switch," Proc. IEEE INFOCOM '96, March 1996, pp. 296-302, describes several algorithms based on weighted maximum bipartite matching (defined therein) and capable of providing 100% throughput in an input-buffered switch. Unfortunately, the  
15 complexity of these algorithms is viewed as too high to be realistic for high-speed hardware implementations. In addition, the nature of the delay guarantees provided by these algorithms remains largely unknown.

Published research by D. Stiliadis and A. Varma, entitled "Providing Bandwidth Guarantees in an Input-Buffered Crossbar Switch," Proc. IEEE INFOCOM '95, April 1995, pp.  
20 960-968, suggests that bandwidth guarantees in an input buffered crossbar switch may be realized using an algorithm referred to as Weighted Probabilistic Iterative Matching (WPIM), which is essentially a weighted version of the algorithm described in Anderson et al. Although the WPIM algorithm is more suitable for hardware implementations than that described by McKeown et al., it does not appear to provide bandwidth guarantees.

One known method of providing bandwidth and delay guarantees in an input-buffered crossbar architecture uses statically computed schedule tables (an example of which is described  
25 in Anderson et al.); however, there are several significant limitations associated with this approach. First, the computation of schedule tables is extremely complex and time-consuming. Therefore, it can only be performed at connection setup-time. Adding a new flow or changing  
30 the rates of the existing flows is quite difficult and time-consuming, since such modifications can require re-computation of the whole table. Without such re-computation, it is frequently

impossible to provide delay and even bandwidth guarantees even for a feasible rate assignment. Consequently, these table updates tend to be performed less frequently than may be desired. Second, per-packet delay guarantees of the existing flows can be temporarily violated due to such re-computation. Third, there exists the necessity to constrain the supported rates to a rather coarse rate granularity and to restrict the smallest supported rate in order to limit the size of the schedule table. All of these limitations serve to substantially reduce the flexibility of providing QoS.

At this time, no other algorithms for providing bandwidth and delay guarantees in input-buffered crossbars are known to the inventors hereof. In one approach, an algorithm allows for 10 the emulation of a non-blocking output-buffered switch with an output FIFO queue by using an input-buffered crossbar with speedup independent of the size of the switch. See B. Prabhakar and N. McKeown, "On the Speedup Required for Combined Input and Output Queued Switching," Computer Systems Lab. Technical Report CSL-TR-97-738, Stanford University. More specifically, this reference shows that such emulation is possible with a speedup of four 15 and conjectures that a speedup of two may suffice. This result allows one to emulate a particular instantiation of a non-blocking output-buffered architecture without having to use the speedup of the order of the switch size (i.e., speedup equal to the number of ports). However, this algorithm is only capable of a very limited emulation of an output buffered switch with FIFO service. Furthermore, as described in the above-referenced technical report, such emulation does 20 not provide any delay guarantees. Its capability of providing bandwidth guarantees over a large time scale is limited to flows which are already shaped according to their rate at the input to the switch, and no bandwidth guarantees can be provided in the presence of misbehaving flows.

It should be noted that in speeded-up input buffered architectures, the instantaneous rate of data entering an output channel may exceed the channel capacity. Therefore, buffering is 25 required not only at the inputs, but also at the outputs. Therefore, input-buffered crossbar switches with speedup are also known as combined input/output buffered switches. Hereinafter, the more conventional term "speeded-up input-buffered crossbar" shall be used.

Another published study of speeded-up input buffered switches suggests that input-buffered switches with even small values of speedup may be capable of providing delays 30 comparable to those of output-buffered switches, but is silent as to the kind (if any) of worst case guarantees provided in the framework described therein. See R. Guerin and K. Sivarajan, "Delay

and Throughput Performance of Speeded-up Input-Queuing Packet Switches," IBM Research Report RC 20892, June 1997.

Thus, there exists a present need in the art to provide deterministic delay and bandwidth guarantees while utilizing the scalability of a crossbar architecture with speedup.

5

### SUMMARY OF THE INVENTION

Accordingly, it is an object of the present invention to provide deterministic delay and bandwidth guarantees in an input-buffered switch with speedup.

It is yet another object of the present invention to ensure the bandwidth and delay guarantees to all flows independently of the behavior of other flows.

It is still yet another object of the present invention to accommodate dynamically changing load and flow composition while operating at high speed, as well as avoid the imposition of artificial restrictions on supported rates.

In accordance with the purposes of the present invention, as embodied and described herein, the above and other purposes are attained by an apparatus and corresponding method for providing bandwidth and delay guarantees in an input-buffered switch with speed-up having input channels and output channels for transferring cells therebetween. The apparatus includes a set of flow queues corresponding to individual flows and a set of per-output-channel queues in each of the input channels for buffering cells awaiting transfer to the output channels. Also residing in each input channel is a flow-level scheduler for scheduling the flow queues and assigning cells for such flow queues to the appropriate per-output-channel queues. Each flow has a rate assigned to it (for example, at connection setup time, or during renegotiation during the lifetime of a flow). Each per-output-channel queue, which corresponds to a different output channel, has an assigned rate and an ideal service associated therewith. The rate of a per-output-channel queue is computed as the sum of rates of all flows destined from the input channel to the output channel corresponding to this queue.

Also included in the apparatus of the present invention is a rate controller corresponding to each input channel for operating on the per-output-channel queues. The rate controller schedules, for a given cell slot, the per-output-channel queues in the input channel to which it corresponds. The rate controller is defined so as to guarantee to each corresponding per-output-channel queue an amount of actual service that is within fixed bounds from the ideal service of

that per-channel-output queue. The fixed bounds, which are constants expressed in units of cells, each are equal to one cell. There is further included an arbiter, responsive to the scheduling of each per-output-channel queue by the rate controller to which it corresponds, for controlling the processing of the queued cells in the scheduled queues through the switch from the input channels 5 to the output channels at a speedup S equal to a number of phases per cell slot. The arbiter, which associates with each scheduled queue a timestamp equal to the time at which such queue was scheduled, uses a specific maximal match computation to choose ones of the scheduled per-output-channel queues from which a cell may be transmitted in each phase.

Each rate scheduler runs a Rate-controlled Smallest Eligible Finish Time First (RSEFTF) 10 algorithm. For each per-output-channel queue, the rate scheduler maintains a first and a second state variable, the first state variable corresponding to an ideal beginning time of the next cell of the per-output-channel queue and the second state variable corresponding to an ideal finishing time of transmission of the next cell of the per-output-channel queue. The rate controller selects as eligible all per-output-channel queues having an ideal beginning time which is less than or 15 equal to a current channel clock counter value. It then chooses as scheduled the eligible queue having the smallest finish time and, for the chosen eligible queue, updates the first state variable with the ideal finish time and second state variable with the ideal beginning time plus one divided by the assigned rate of the queue.

According to another aspect of the present invention, the specific maximal match 20 computation performed by the arbiter is as follows. First, there is provided a Set\_Match set and Set\_Queues set, the former being initialized to an empty set and the latter to the set of the associated timestamps. Next, the arbiter selects the smallest of the associated timestamps stored in Set\_Queues, consequently adding the selected associated timestamp to Set\_Match and removing the selected associated timestamp from Set\_Queues. All remaining associated 25 timestamps associated with per-output-channel queues corresponding to either the same input channel or output channel as the selected associated timestamp are then deleted from Set\_Queues. If the Set\_Queues set is empty, the arbiter sends an indication of the per-output-channel queues corresponding to the timestamps in the Set\_Match set to the input channels to which they belong. If the Set\_Queues set is not empty, the computation returns to the step of selecting.

30 In an alternative embodiment of the present invention, the queue scheduling of the rate controller is a centralized function. That is, the arbiter runs all of the rate controllers locally.

The present invention achieves several important goals. It provides deterministic delay guarantees comparable to delay guarantees associated with non-blocking output-buffered architectures, while utilizing the scalability of a crossbar architecture. It allows arbitrary assignment of guaranteed rates (as long as the rates are feasible in the sense that the sum of all 5 rates does not exceed the total available bandwidth at any input or any output). Additionally, it allows the flexibility to quickly admit new flows and change the rate assignment of existing flows. Moreover, it provides the protection of these guarantees to well-behaved flows even in the presence of misbehaved flows.

More specifically, it can be proved that deterministic bandwidth and delay guarantees can 10 be obtained with speedup  $S$  greater than or equal to 3 in a switch with 100% load of all links. With no speedup or with small values of speedup, the system is capable of providing similar deterministic guarantees if the load due to guaranteed flows is limited to a certain portion ( $S/3$ ) of the bandwidth of any link. The remaining bandwidth can be used by best-effort traffic. The proof of these statements can be found in the Appendix. Furthermore, it is conjectured that 15 similar guarantees can be provided with speedup of greater than or equal to 2 and full load of the link, and with any speedup greater than or equal to one with load due to guaranteed flows limited to 50% of the link bandwidth.

While the invention is primarily related to providing bandwidth and delay guarantees to 20 flows requiring such guarantees, it can be used in conjunction with best-effort traffic which does not require such guarantees. If best effort traffic is present, it is assumed that the invention as described herein is run at an absolute priority over any scheduling algorithm for best effort traffic.

#### BRIEF DESCRIPTION OF THE DRAWINGS

The above objects, features and advantages of the present invention will become more 25 apparent from the following description of the embodiments of the present invention illustrated in the accompanying drawings, wherein:

Fig. 1 is block diagram depicting an input-buffered crossbar switch capable of utilizing per-output-channel queue scheduling and arbitration schemes in accordance with the present invention;

30 Fig. 2 is a flow diagram illustrating the actions of the input channel related to scheduling per-output-channel queues;

Fig. 3 is a flow diagram depicting the arbitration policy for providing bandwidth and delay guarantees in accordance with the present invention; and

Fig. 4 is a depiction of one example of the scheduling policy for per-output-channel queues in the input channel shown in Fig. 2.

5

### DETAILED DESCRIPTION

Referring to FIG. 1, with like reference numerals identifying like elements, there is shown an input-buffered crossbar switch 10 implementing a crossbar arbitration scheme in accordance with the present invention. As illustrated in FIG. 1, the underlying architecture of the input-buffered crossbar switch 10 is represented as an  $n \times m$  crossbar. Here, " $n$ " is the number of input channels  $i$  ( $1 \leq i \leq n$ ) 12 and " $m$ " is the number of output channels  $j$  ( $1 \leq j \leq m$ ) 14. Each input channel has one or more input ports 16, each of which corresponds to a physical input link 18. Similarly, the output channels each have one or more output ports 20, each corresponding to a physical output link 22. The input channels 12 are connected to the output channels 14 by way of a crossbar unit 24. It will be understood by those skilled in the art that the crossbar unit as depicted in FIG. 1 includes a crossbar switch fabric of known construction, the details of which have been omitted for purposes of simplification. It is the crossbar switch fabric that is responsible for transferring cells between input and output channels.

In the embodiment shown, the total capacity of all input channels and all output channels is assumed to be the same, although the capacity of individual links may be different. Hereinafter, the capacity of a single channel is denoted by  $r_c$ . The speed of the switch fabric, denoted by  $r_{sw}$ , is assumed to be  $S$  times faster than the speed of any channel. In general, the switch and the channel clocks are not assumed to be synchronized. The speedup values may be arbitrary (and not necessarily integer) values in the range of  $1 \leq S \leq n$ . It is further assumed that the switch operates in phases of duration  $T_{sw}$  defined as the time needed to transmit a unit of data at speed  $r_{sw}$ . Such phases are referred to as matching phases. In this disclosure, a unit of data shall be referred to as a *cell*. Accordingly, a switch can move at most one cell from each input channel and at most one cell to each output channel at each matching phase. Therefore, on the average, a switch with speedup  $S$  can move  $S$  cells from each input channel and  $S$  cells to each output channel. At  $S=n$ , the switch is equivalent to the output buffered switch.

Although not shown in FIG. 1, packets received on a given input link 18 are typically

buffered at the input ports. Also, each flow to which the received packets correspond may be allocated a separate buffer or queue at the input channel. These "per-flow" queues may be located in an area of central memory within the input channel. Alternatively, flow queues may be located in a memory in the input ports associated with the input channel. When the packets received from the input links are of variable length, they are fragmented into fixed-size cells. If the packets arriving at the switch all have a fixed length (e.g. cell in ATM networks), no fragmentation is required. In packet switching networks, where arriving packets are of different size, the implementation is free to choose the size of the cell as convenient. The tradeoff in the choice of this size is that the smaller the cell, the better delay guarantees can be provided, but the faster the switch fabric must be (and therefore the more expensive the switch). Small cell size also increases fragmentation overhead. Upon arrival and after possible fragmentation, cells are mapped to a corresponding flow (based on various classifiers: source address, destination address, protocol type, etc.). Once mapped, the cells are placed in the appropriate "per-flow" queue.

Associated with each flow requiring bandwidth and/or delay guarantees is some rate  $r_f$ . Typically, for guaranteed rate or guaranteed delay flows, the rate  $r_f$  is established at connection setup time (e.g., via RSVP). Rates assigned to guaranteed flows can also be changed during a renegotiation of service parameters as allowed by the current RSVP specification. It is assumed that the rate assignment is feasible, i.e., the sum of the rates of all flows at each input port does not exceed the capacity of this input port, and the sum of rates of all flows across all input ports destined to a particular output port does not exceed the capacity of that output port. The feasibility of rates across all input and output ports implies the feasibility of rates across all input and output channels. Included in the rate  $r_f$  guaranteed to the flow is any overhead associated with packet fragmentation and re-assembly. The actual data rate negotiated at connection setup may therefore be lower. For networks with fixed packet size, such as ATM, however, no segmentation and re-assembly is required. Thus, no overhead is present.

As shown in FIG. 1, each input channel  $i$  has  $m$  virtual output queues (VOQs) or per-output-channel queues 26 (also referred to as per-output or virtual output queues), denoted by  $Q(i,j)$ ,  $1 \leq j \leq m$ , one for each output channel  $j$  14. In the embodiment shown in FIG. 1, the input channel maintains a single flow-level scheduler  $S_f(i)$  28, which needs to schedule only a single flow per cell time. Once scheduler  $S_f(i)$  schedules some flow  $f$ , it adds the index of this

flow  $f$  (or, alternatively, the head of the line (HOL) cell of flow  $f$ ) to the tail of queue  $Q(i,j)$ . Thus, depending on the implementation,  $Q(i,j)$  may contain either cells or pointers to cells of individual flows.

Also located at each input channel 12 is a rate controller 30 (or scheduler  $S_q(i)$ ), which  
5 schedules or selects for processing one (or more) of the virtual output queues  $Q(i,j)$  at every channel cell time. The channel cell time is defined as the time required to transmit one cell at channel speed. The scheduler  $S_q$  ensures that the aggregate of all flows destined from a given input to a given output is guaranteed the bandwidth  $r_{i,j}$  equal to the sum of individual flow rate assignment over all flows destined from a given input to a given output, as well as per-cell delay.  
10 An example of scheduler  $S_q$  is described in paragraphs below. Indices of (or pointers to) the queues  $Q(i,j)$  chosen at each cell time are given to an arbiter 32, which is located in the crossbar unit 24. It is the arbiter's responsibility to determine which of the input channels should be able to transmit a cell to particular output channels. It is assumed that the arbiter operates in phases, also referred to as "matching phases". The duration of each phase is equal to the duration of the  
15 channel cell slot divided by the speedup  $S$ . The goal of the arbiter is to compute a maximal (conflict-free) match between the input and output channels so that at most one cell leaves any input channel and at most one cell enters any output channel during a single matching phase. Although the term "maximal match" (or, alternatively, "maximal matching") is well understood by those skilled in the art, a definition may be had with reference to papers by N. McKeown et al. and Stiliadis et al., cited above, as well as U.S. Patent No. 5,517,495 to Lund et al.  
20

The arbiter maintains  $n \times m$  queues 34, denoted by  $A(i,j)$ , with each arbiter queue corresponding to a different one of  $Q(i,j)$ . The arbiter queues 34 are used to store timestamps as described below. At each channel cell time the arbiter receives the index (or indices) of the per-output queue(s)  $Q(i,j)$  chosen by the input scheduler  $S_q(i)$  at this channel slot time. When the arbiter 32 receives the index of some  $Q(i,j)$ , it adds a timestamp equal to the current time into the corresponding queue  $A(i,j)$ .

As explained above, during each of its matching phases, the arbiter decides which input can send a cell to which output by computing a maximal matching between all inputs and all outputs. The algorithm used to compute the maximal match is described in detail in paragraphs  
30 to follow. Once the matching is completed, the arbiter notifies each input of the output to which it can send a cell by sending to the input channel the index of the per-output queue from which

the cell is to be transmitted. The input channel then picks a cell to send to that output channel and the cell is transmitted to the output channel.

When an input channel 12 receives from the arbiter the index of the  $Q(i,j)$  corresponding to the output channel for the current matching phase, it forwards the HOL cell of  $Q(i,j)$  (or, 5 alternatively, the cell pointed to by the HOL pointer in  $Q(i,j)$ ) to the output channel  $j$ . If  $Q(i,j)$  is empty (that is, there is no cell of a guaranteed flow in the queue), then a cell of a lower-priority service destined to the same output is sent instead. If there is no best effort traffic at this input matching phase, then no cell is sent. The size of  $Q(i,j)$  is determined by the properties of the schedulers  $S_f$  and  $S_q$ .

10 In another variation, each input channel could maintain one flow-level scheduler  $S_f(i,j)$  for each output. When the input channel  $i$  needs to transmit a cell to a given output  $j$ , it invokes scheduler  $S_f(i,j)$  to determine which flow destined to  $j$  should be chosen. Unlike the option described above, in which scheduler  $S_f(i)$  can run at link speed, the flow-level schedulers  $S_f(i,j)$  must be capable of choosing up to  $S$  cells per cell time as it is possible that this input may need 15 to send a cell to the same output in all  $S$  matching phases of the current cell slot. In yet another approach, the input can run  $m$  parallel  $S_f$  schedulers, one per output. Each of these schedulers may schedule  $1 \leq k \leq S$  cells per cell time. When a flow is scheduled by  $S_f$ , an index to this flow is added to  $Q(i,j)$ .

Although not shown in FIG. 1, a cell forwarded by an input channel  $i$  to an output channel 20  $j$  is added to a queue maintained by the output channel. A variety of queuing disciplines can be used, such as FIFO, per-input-port, or per flow. If the queue is not a simple FIFO, each output has an additional scheduler, shown in FIG. 1 as output scheduler  $S_o$  36. This output scheduler determines the order in which cells are transmitted onto the output link from the output channel. It is assumed that any required reassembly occurs before  $S_o$  is used, so that  $S_o$  schedules 25 packets rather than cells.

Any known QoS-capable scheduler such as those mentioned can be used for the scheduler  $S_f$  and  $S_o$ .

Since each scheduler  $S_f$ ,  $S_o$  operates independently of the other, the delay of an individual cell in the switch is the sum of the delay of this cell under its input and output 30 schedulers  $S_f$  and  $S_o$ , plus the delay due to the potential arbitration conflicts. The delay of a packet segmented in cells is comprised of the delay experienced by its last cell plus the

segmentation and re-assembly delays.

Still referring to FIG. 1, it can now be appreciated that, with respect to each input channel,  $S_q$  operates on each of the queues  $Q(i,j)$  VOQ containing cells (or pointers to cells) which have already been scheduled by  $S_f$  but which have not yet been transmitted to their destination output channel with which the VOQ is associated due to arbitration conflicts. The present invention undertakes the task of determining the sequence of transmissions between input channels and output channels satisfying the crossbar constraint that only one cell can leave an input channel and enter an output channel per phase in such a way that the arbitration delay is bounded for each cell awaiting its transmission at the input channel. In the system as partitioned in the embodiment shown in FIG. 1, this task is distributed among the arbiter and the input channels. This task is discussed in further detail below.

Now referring to FIG. 2, there is illustrated the actions of the input channel with respect to scheduling the per-output channel queues and the interaction with the arbiter an arbitration policy 40 in accordance with the present invention. At the initial step 42, the input channel initializes associated state variables and obtains the assigned rates  $r_{i,j}$  (again, equal to the sum of rates of all flows going from input  $i$  to output  $j$ ) for the queues  $Q(i,j)$ . The sum of all rates is feasible in that the sum of rates  $r_{i,j}$  across all inputs channels  $i$  is less than or equal to the channel rate  $r_c$  and the sum of rates  $r_{i,j}$  across all outputs channels  $j$  is less than or equal to  $r_c$ . The feasibility is ensured by admission control or by other means not discussed herein. In step 20 44 the input channel initializes its clock counter (denoted *time*) to zero. The unit of time for this clock is one channel cell slot. Further, in step 46, which is invoked each cell slot, the scheduler  $S_q$  is run to determine the queue  $Q(i,j)$  to be selected or scheduled during this time slot. The operation of scheduler  $S_q$  will be discussed in more detail later with reference to FIG. 4. In step 48 the index of the queue selected in step 46 is passed to the arbiter. In step 50 the input channel checks if the arbiter has notified it of any of its queues  $Q(i,j)$  having been matched during this cell slot. If so, the HOL cell in the matched queue is transmitted to the corresponding output channel in step 52. Steps 50, 52, 54 may be repeated several times during one cell slot, since there may be several matching phases occurring during one cell slot. At the end of the cell slot the clock counter is incremented (step 56). To accommodate potential rate changes the rates  $r_{i,j}$  30 are updated if necessary as indicated in step 58.

The operation of the arbiter is now described with reference to FIG. 3. As illustrated in

FIG. 3 at the initial step 60, the arbiter clock is initialized to zero. The unit of time for this clock is the duration of one matching phase. At the beginning of any matching phase, the arbiter checks if any queue indices have been sent to it by input channels (step 62). If the arbiter has not received any queue indices, it determines when the current matching phase is over at step 64.

5 Once it has been determined that the current phase is over, the clock counter *time* is incremented by one at step 66. If the arbiter has received any queue indices in the current phase, it adds a timestamp equal to the current clock counter value to the tail of the queues  $A(i,j)$  corresponding to the queues  $Q(i,j)$  whose indices have been received in step 62 (step 68).

The maximal match computation 69, which is used to generate a match, is performed according

10 to the present invention as shown in steps 70, 72, 74, 76, 78. A match is defined as a conflict-free set of input/output pairs and is computed based on the current contents of queues  $A(i,j)$  at the arbiter. Beginning with step 70, the arbiter initializes a first set *Set\_Match* to an empty set and a second set *Set\_Queues* to the set of all non-empty queues  $A(i,j)$ . In step 72, the arbiter finds the smallest timestamp of all Head-of-Line (HOL) timestamps in set *Set\_Queues*. The queue

15  $A(i,j)$  containing the smallest timestamp found in step 72 is added to *Set\_Match* in step 74. In step 76, the arbiter removes from set *Set\_Queues* all queues  $A(i,k)$  and  $A(k,j)$  where  $i$  and  $j$  are the input and output of the queue  $A(i,j)$  selected in step 72 (these are all queues corresponding to the same input and the same output as those of  $A(i,j)$  already added to the match). If there are no more unmatched queues (step 78), the match is complete. Consequently, the arbiter sends the

20 inputs the indices of the matched queues at step 80 and returns to step 66. Otherwise, the arbiter goes to the next iteration of the matching process (by returning to step 72).

Another essential element of the invention is the choice of a rate-controller  $S_q$  operating on the per-output queues  $Q(i,j)$  at each input. More specifically, rate-controller  $S_q$  is defined as a rate controller capable of guaranteeing each per-output queue its assigned rate  $r_{i,j}$  and offering an amount of service  $W_{i,j}(t)$  to each per-output queue in any interval  $(0,t)$  so as to satisfy the following condition

$$tr_{i,j} - E1 \leq W_{i,j}(t) \leq tr_{i,j} + E2 \quad (\text{eq1})$$

where  $E1$  and  $E2$  are the early and late and early work discrepancy bounds, respectively, which are assumed to be constants and are expressed in the units of cells.  $E1$  and  $E2$  should be interpreted as the lower and upper bounds on the discrepancy between  $W_{i,j}(t)$  (the amount of actual service given by  $S_q$  to queue  $Q(i,j)$ , expressed in cells) and the ideal service  $tr_{i,j}$  cells

that same queue should have received in this interval under the fluid model (recall that the rates are assumed to be expressed in units of cells per a unit time). Thus, any  $S_q$  satisfying the above condition at all times  $t$  can be used to provide rate and bandwidth guarantees with the values of speedup independent of the size of the switch.

One instantiation of a scheduler  $S_q$  which performs the queue scheduling operation depicted in step 46 of FIG. 2 and satisfies the above-stated condition (eq1) is now considered with reference to FIG. 4. This particular rate controller, essentially equivalent to a rate-controlled version of WF2Q by Bennett et. al., mentioned above, is based on a Rate-controlled Smallest Eligible Finish Time First (RSEFTF) algorithm. For each of the per-output queues  $Q(i,j)$ , RSEFTF the rate controller maintains two state variables: a first state variable  $b_{ij}$ , which is the ideal beginning time of transmission of the next cell of this queue, and a second state variable  $f_{ij}$ , which is the ideal finishing time of transmission of the next cell of this queue. Initially,  $b_{ij} = 0$  and  $f_{ij} = 1/r_{ij}$  (the initialization of these state variables occurs in step 42 of FIG. 2). As shown in step 80 of FIG. 4, the scheduler selects all queues at this input for which  $b_{ij}$  is less than or equal to the current channel cell slot time (channel clock counter *time*). Such queues are called eligible queues. In step 82, the scheduler then chooses as scheduled the queue  $Q(i,j)$  with the smallest finish time  $f_{ij}$  from among the eligible queues. Ties are broken arbitrarily. The queue chosen in step 82 is the one whose index is sent to the arbiter in step 48 of FIG. 2. In step 84, the scheduler updates the state variables  $b_{ij}$  and  $f_{ij}$  corresponding to the chosen queue as follows:  $b_{ij} \leftarrow f_{ij}$ ,  $f_{ij} \leftarrow b_{ij} + 1/r_{ij}$ . The variables  $b_{ij'}$  and  $f_{ij'}$  for all  $j'$  not equal to  $j$ , i.e., for the queues not chosen at the current cell slot, remain unchanged.

It can be shown that the described embodiment satisfies several properties. Property 1 is as follows: In an  $n \times m$  crossbar running RSEFTF with integer speedup  $S \geq 3$  and matching phases synchronized with input cell times, arbitration delay of any cell never exceeds  $n$  channel cell times. More specifically, any cell scheduled at the beginning of cell slot  $t$  is transmitted before the beginning of cell slot  $t+n$ . This is true for arbitrary feasible rate assignment. Property 2 is as follows: In an  $n \times m$  crossbar running RSEFTF, for arbitrary (not necessarily integer) speedup  $S \geq 3$  and no assumption on synchronization between the channel clock and the switch clock, i.e., no alignment of matching phases and cell slots is assumed, arbitration delay of any cell is bounded by  $n$  as long as for any output  $j$  the rate assignment satisfies  $S_i(r_{ij}) < 1$ . The proof of properties 1 and 2 is given in the Appendix. Note that Property 1 holds for any feasible rate assignment, while Property 2 has been proved only for the case when the sum of rates of all

queues at all inputs corresponding to an output channel is strictly less than the capacity of the output channel. However, since this sum can be arbitrarily close to the channel capacity, for all practical purposes this limitation is unimportant. Although properties 1 and 2 have been proven only for speedup  $S \geq 3$ , simulations indicate that the speedup  $S \geq 2$  is sufficient to obtain the delay bound of  $n$ . It can be conjectured, therefore, that Properties 1 and 2 also hold for  $S=2$ .

Since the cost of the switch is typically higher the larger the speedup, the application of the present invention in switches with speedup  $1 \leq S \leq 3$  is now explored in some detail. Here, the method and apparatus of the present invention can be used without modification if, instead of allowing arbitrary feasible rate assignment, the total bandwidth allocated to guaranteed flows is restricted to a certain portion of the link bandwidth. More specifically, it can be shown that: for arbitrary  $1 \leq S < 3$ , with no assumption on synchronization of cell slot and phase clocks, if the rate assignment of guaranteed flows satisfies (1)  $S_i(r_{i,j}) < S/3$ ,  $S_j(r_{i,j}) \leq S/3$  OR (2)  $S_i(r_{i,j}) \leq S/3$ ,  $S_j(r_{i,j}) < S/3$ , then the arbitration delay of any cell is bounded by  $3(n+m)/S - 1$  [Property 3]. The proof of Property 3 is also given in the Appendix. Property 3 demonstrates that even in switches with small speedup or no speedup at all, as long as the sum of rates of guaranteed flows does not exceed the ratio  $S/3$  of the link bandwidth, each cell can be guaranteed deterministic arbitration delay and therefore deterministic total delay in the switch. The restriction on the sum of the rates can be achieved by admission control or by other means, which are not discussed in this disclosure.

There are many ways to implement various known components of the described embodiment. Therefore, the details of such implementations are largely omitted from the disclosure. However, memory requirements at the arbiter are now discussed briefly. The arbiter needs to store timestamps of the cells scheduled by input schedulers  $S_q$  but not yet transmitted to their outputs. Memory requirements depend on the size of the timestamps stored in the arbiter and the number of such timestamps. Note first that the timestamps are integers, since they represent the matching phase number. At first glance, it may appear that the timestamps can grow infinitely large as the execution time increases. This would be impractical since the size of the timestamps and, consequently, memory requirements would be unbounded. However, the fact that the arbitration delay is bounded means that the difference between any two timestamps is bounded as well. This allows for the reduction in the range of timestamps to a limited range. The bound on arbitration delay also implies that the number of the timestamps that can be stored in the arbiter is limited as well. For example, Properties 1 and 2 demonstrate that with speedup

$S \geq 3$  the arbitration delay of any cell can be at most  $n$ . Since for each input  $i$  at most one cell is scheduled per channel slot, it is clear that there can not be more than  $n$  timestamps in all queues  $A(i,j)$  corresponding to a single input  $i$ . It can be shown that this implies that no queue can contain more than  $n+1$  timestamps, which implies that the total number of timestamps in the arbiter cannot exceed  $n(n+1)$ . Therefore, at an additional expense of storage required for accounting to which queue a timestamp belongs, there is no need to store more than  $n(n+1)$  timestamps, since the memory can be dynamically allocated to a newly arrived timestamp. If the implementation chooses to statically allocate memory for all  $A(i,j)$ , however, then the total amount of memory required is  $mn(n+1)$ , because in the worst case  $n+1$  timestamps can be in any of the  $n \times m$  queues.

Alternatively, the arbiter need not maintain all the timestamps of all scheduled cells. Instead, part of the burden of maintaining timestamps is shifted to the input channel. This can be explained more clearly with reference to steps 486 of FIG. 2 and step 468 of FIG. 3. At step 48, the timestamps communicated by the input channel to the arbiter at the beginning of each matching phase include only those timestamps corresponding to scheduled HOL cells in each queue. If the scheduler  $S_q$  can only schedule one cell per cell time (that is, the scheduler  $S_q$  operates at the speed of the input channel), then the input channel need only communicate at most two timestamps to the arbiter: one for the queue whose cell was transmitted to the output at the end of the previous matching phase (and there could only be at most one such cell per input), and one for the cell which has just been scheduled by  $S_q$  at the current channel cell time. If the scheduler  $S_q$  operates faster than the speed of the channel, then more than one cell can be scheduled per channel cell time. At step 68, then, the arbiter receives and at most two timestamps per input/output pair. Consequently, since only one timestamp now needs to be stored per each  $A(i,j)$ , memory requirements of the arbiter are reduced, but at the expense of increased communications between the input channels and the arbiter and at the expense of additional storage at the input, as the input channel now needs to store not only the cell (or the pointer to the cell) scheduled by  $S_q$ , but also the time at which it was scheduled.

As described above and shown in FIG. 1, the  $S_q$  portion of the arbitration mechanism is decentralized at the input channels. Alternatively, the arbiter can run all of the rate controllers  $S_q$  locally. To do so, the arbiter treats the arrivals to the per-output queues at the input channels as occurring exactly at their ideal times and at their ideal rates. Conceptually, it maintains for each input/output pair a queue  $A(i,j)$  to which imaginary "dummy" cells arrive according to their

ideal inter-arrival times. It then runs  $n$  copies of the  $S_q$  rate controller (one per each input) to determine the scheduling order of these dummy cells. The operation of the input channel is now simplified by omitting steps 46 and 48 of FIG. 2. While step 48 is eliminated completely, step 46 as performed by the arbiter for all input channels in parallel (step 46') replaces steps 62, 64 and 66 in FIG. 3. The rest of the operation follows that of the distributed version described above in reference to FIG. 2 and FIG. 3.

As should be clear from this description, the arbiter is not provided with timestamps or queue indices for queues scheduled by each input channel at each cell time. Rather, the arbiter obtains this information for itself. Unlike the implementations previously described, this approach clearly puts a higher load on the arbiter in terms of computation and storage, while reducing the amount of communication between the input channels and the arbiter. With this approach, it is possible that some queue  $Q(i,j)$  at the input channel is matched when this queue is empty. In this case the input channel simply does not send any guaranteed cell. If best effort traffic is present, the opportunity to transmit the cell is passed to best effort traffic.

In all options of the preferred embodiment as described thus far, there is a scheduler  $S_q$  per each input channel. Where the schedulers  $S_q$  are run by the input channels, the computational work required for the scheduling decision is distributed among input channels and therefore the computational load is reduced. As discussed above, it is also possible to perform all scheduling decisions at the arbiter by running in parallel schedulers  $S_q$  for each of the input channels. In this case, the complexity of the arbiter is increased, but there is substantial saving in the amount of communication required between input channels and the arbiter. Properties 1,2 and 3 described above hold for either one of these two options.

While the disclosed input-buffered switch and scheduling method has been particularly shown and described with reference to the preferred embodiments, it will be understood by those skilled in the art that various modifications in form and detail may be made therein without departing from the scope of the invention as set forth by the claims. Accordingly, modifications such as those suggested above, but not limited thereto, are to be considered within the scope of the claims.

## APPENDIX

It is assumed that all cells are scheduled by their  $S_q$  scheduler at the beginning of a channel cell slot and immediately become available for the arbiter. Matching occurs at the right-hand boundary, i.e., at the end, of a matching phase. Only cells available to the arbiter at the 5 beginning of a matching phase participate in the matching process during this phase. It is also assumed that, if a phase boundary coincides with a slot boundary, all departures at this phase boundary occur prior to the scheduling decisions of  $S_q$  at the beginning of the next cell slot. The term "cell slot" shall be used to denote channel cell slot.

**Lemma 1.** If a cell  $c(i,j)$  has been scheduled before or at the beginning of a matching phase and 10 remains there at the end of this matching phase, then a cell with smaller or equal scheduling time went from input  $i$  or to output  $j$ .

### **Proof.**

By operation of the algorithm matching occurs in the increasing order of scheduling times. Therefore, if no other cell has gone from input  $i$  or to output  $j$  in a match, by operation of the 15 algorithm, this cell would have gone. This is a simple consequence of the input-output contention.

**Definition.** Arbitration delay of a cell is the time elapsed between the time it is scheduled by its input scheduler  $S_q$  and the time it is transferred to its output.

**Property 1.** In an  $n \times m$  crossbar ( $n$  inputs  $m$  outputs) running RSEFTF with integer speedup 20  $S \geq 3$  and matching phases aligned with input cell times, arbitration delay of any cell never exceeds  $n$  channel cell times. More specifically, any cell scheduled at the beginning of cell slot  $t$  is transmitted before the beginning of cell slot  $t+n$ . This is true for arbitrary feasible rate assignment.

### **Proof.**

25 Consider the first time  $t$  this is violated for one or more cells in the switch. There must be one or more cells which were scheduled at or before the beginning of cell slot  $t-n$  and which still remain at their inputs at the beginning of cell slot  $t$ . Let  $c(i,j)$  denote the cell with the earliest scheduling time of all such cells (if there are several cells with the same earliest scheduling time, we pick any of them). This cell must have been scheduled by its scheduler  $S_q$  at the beginning 30 of cell slot  $t-n$ . This is because if it were scheduled later, it would not have violated its delay bound  $n$  by time  $t$ , and if it were scheduled earlier, its delay bound would have been violated before  $t$ , which would contradict the assumption that  $t$  was the first time a violation

occurred. For integer speedup  $S$  with matching phases synchronized with cell times, there are exactly  $nS$  matching phases which had occurred between the beginning of cell slot  $t-n$  and the beginning of cell slot  $t$ . By Lemma 1, in order for cells  $c(i,j)$  to remain at its input at time  $t$ , there must have been at least  $nS$  cells with smaller or equal timestamps than  $t-n$  (which is the timestamp of  $c(i,j)$ ) which were scheduled by the  $S_q$  scheduler at input  $i$  and/or by any input scheduler  $S_q$  destined to output  $j$ .

Clearly, no cell scheduled after time  $t-n$  will have a timestamp less than or equal to  $t-n$ . This means that, in order for  $c(i,j)$  to remain at its input at time  $t$ , at time  $t-n$  there must have been at least  $nS$  cells aside from  $c(i,j)$  at input  $i$  and/or at any input destined to output  $j$ , which had 10 been scheduled by the scheduler  $S_q$  at or prior to the beginning of time  $t-n$ , but not transmitted yet by that time. By the assumption of  $t-n$  being the earliest scheduling time of any cell which violated delay bound  $n$ , all cells which were scheduled by their  $S_q$  schedulers at or prior to time  $t-2n$  would have been transmitted by the beginning of cell slot  $t-n$ . Therefore, the only cells which can be not yet transmitted by time  $t-n$  are those which were scheduled in  $n$  cell slots  $t-15 2n+1, t-2n+2 \dots t-n$ , which includes  $c(i,j)$  itself. Since the input scheduler  $S_q$  schedules at most one cell per cell slot, there can be at most  $n-1$  such cells at input  $i$  (not counting  $c(i,j)$ ). By the properties of RSEFTF, each queue  $(i',j')$  can be scheduled at most  $n r_{i',j'} + 1$  times in any  $n$  cell slots. Therefore, there can be at most  $S_i(n r_{i,j}) + n-1$  cells scheduled in the interval  $(t-n-(n-1), t-n)$  of  $n$  cells slots long to output  $j$  (not counting  $c(i,j)$ ). Therefore, the total number of cells at 20 time  $t-n$  which are either at input  $i$  or destined to output  $j$  which could still be at their inputs at the beginning of cell slot  $t-n$  with time stamps less than or equal to  $t-n$ , not counting  $c(i,j)$  is at most  $2n + S_i(n r_{i,j}) - 1 < 3n$ . Therefore, for any  $S \geq 3$  there will not be enough cells with smaller or equal timestamps than  $c(i,j)$  to prevent  $c(i,j)$  from leaving its input before time  $t$ . QED.

**Property 2.** In an  $n \times m$  crossbar running RSEFTF, for arbitrary (not necessarily integer) speedup  $S \geq 3$ , with no assumption on synchronization between the cell slot clock and the phase clock (no alignment of phases in cell slots), arbitration delay of any cell is bounded by  $n-1$  as long as for any output  $j$   $S_i(r_{i,j}) < 1$ .

#### Proof.

The proof is almost identical to that of Property 1. Consider the first time  $t$  this is violated for 30 one or more cells in the switch. That is, at the beginning of some cell slot  $t$ , there is one or more cells which were scheduled at or before the beginning of cell slot  $t-n$  and which still remain at their inputs. Let  $c(i,j)$  be the cell with the earliest scheduling time of all such cells (if there are

several cells with the same earliest scheduling time, we pick any of them). This cell must have been scheduled by its scheduler  $S_q$  at the beginning of cell slot  $t-n$ . This is because if it were scheduled later, it would not have violated its delay bound  $n$  by time  $t$ , and if it were scheduled earlier, its delay bound would have been violated before  $t$ , which would contradict our 5 assumption that  $t$  was the first time a violation occurred.

It is easy to see that for any speedup  $S$  there are at least  $nS-2$  full matching phases which occur between the beginning of cell slot  $t-n$  and the beginning of cell slot  $t$ . By Lemma 1, in order for cell  $c(i,j)$  to remain at its input at time  $t$ , there must have been at least  $nS-2$  cells with smaller or equal timestamps than  $c(i,j)$  scheduled from  $i$  and/or to  $j$ . Clearly, no cell 10 scheduled after time  $t-n$  will have a timestamp less than or equal than  $t-n$ , which is the timestamp of  $c(i,j)$ . This means that in order for  $c(i,j)$  to remain at its input at time  $t$ , at time  $t-n$  there must have been at least  $nS-2$  cells aside from  $c(i,j)$  at input  $i$  and/or destined to output  $j$  which had been scheduled by the scheduler  $S_q$  at or prior to the beginning of time  $t-n$ , but were not transmitted yet by that time. By the assumption of  $t-n$  being the earliest scheduling time of any 15 cell which violated delay bound  $n$ , all cells which were scheduled prior to time  $t-2n+1$  would have been transmitted by time  $t-n$ . Therefore, the only cells which can be not yet transmitted by time  $t-n$  are those which were scheduled in  $n$  cell slots  $t-2n+1, t-2n+2 \dots t-n$ , which includes  $c(i,j)$  itself. Since the input scheduler  $S_q$  schedules at most one cell per cell slot, there can be at most  $n-1$  such cells (not counting  $c(i,j)$ ) at input  $i$ . By the properties of RSEFTF, each queue  $(i',j')$  can 20 be scheduled at most  $n r_{i',j'} + 1$  times in any  $n$  cell slots. Therefore, there can be strictly less than  $S_i(n r_{i,j}) + n-1$  cells scheduled in the interval  $(t-2n+1, t-n)$  of  $n$  cells slots to output  $j$  (not counting  $c(i,j)$ ). This is because by the statement of the theorem  $S_i(r_{i,j}) < 1$ . Therefore, the total number of cells at time  $t-n$  which are either at input  $i$  or destined to output  $j$  which could still be at their inputs at the beginning of cell slot  $t-n$  with time stamps less than or equal to  $t-n$ , not 25 counting  $c(i,j)$ , is strictly less  $2n + S_i(n r_{i,j}) - 1 < 3n$ . Therefore, for any  $S \geq 3$  there will not be enough cells with smaller or equal timestamps than  $c(i,j)$  to prevent  $c(i,j)$  from leaving its input before time  $t$ . QED.

- Property 3.** For arbitrary  $1 \leq S < 3$ , with no assumption on synchronization of cell slot and phase clocks, if the rate assignment of guaranteed flows satisfies
- 30 (1)  $S_i(r_{i,j}) < S/3, S_j(r_{i,j}) \leq S/3$
- OR
- $S_i(r_{i,j}) \leq S/3, S_j(r_{i,j}) < S/3,$

then the arbitration delay of any cell is bounded by  $3(n+m)/S - 1$

**Proof.**

Consider the first time  $t$  this is violated for one or more cells in the switch. That is, at the beginning of some cell slot  $t$  there is one or more cells which were scheduled at or before the 5 beginning of cell slot  $t-3(n+m)/S$  and which still remain at their inputs. Let  $c(i,j)$  be the cell with the earliest scheduling time of all such cells (if there are several cells with the same earliest scheduling time, we pick any of them). This cell must have been scheduled by its scheduler  $S_q$  at the beginning of cell slot  $t-3(n+m)/S$ . This is because if it were scheduled later, it would not have violated its delay bound by time  $t$ , and if it were scheduled earlier, its delay bound would 10 have been violated before  $t$ , which would contradict the assumption that  $t$  was the first time a violation occurred. It is easy to see that there are at least  $S \cdot 3(n+m)/S - 2 = 3(n+m) - 2$  full matching phases which occur between the beginning of cell slot  $t-3(n+m)/S$  and the beginning of cell slot  $t$ . By Lemma 1 in order for cell  $c(i,j)$  to remain at its input at time  $t$ , there must have been at least  $3(n+m) - 2$  cells with smaller or equal timestamps than  $c(i,j)$  scheduled from  $i$  15 and/or to  $j$ . Clearly, no cell scheduled after time  $t-3(n+m)/S$  will have a timestamp less than or equal than  $t-3(n+m)/S$ , which is the timestamp of  $c(i,j)$ . This means that in order for  $c(i,j)$  to remain at its input at time  $t$ , at time  $t-3(n+m)/S$  there must have been at least  $3(n+m)-2$  cells aside from  $c(i,j)$  at input  $i$  and/or destined to output  $j$  which had been scheduled by the scheduler  $S_q$  at or prior to the beginning of time  $t-3(n+m)/S$ , but were not transmitted yet by that 20 time. By our assumption of  $t-3(n+m)/S$  being the earliest scheduling time of any cell which violated delay bound  $3(n+m) - 1$ , all cells which were scheduled prior to time  $t-3(n+m)/S - (3(n+m)/S - 1)$  would have been transmitted by time  $t-3(n+m)/S$ . Therefore, the only cells which can be not yet transmitted by time  $t-3(n+m)/S$  are those which were scheduled in  $3(n+m)/S$  cell slots  $t-3(n+m)/S - (3(n+m)/S - 1)$ ,  $t-3(n+m)/S - (3(n+m)/S - 2)$ ...  $t-3(n+m)/S$ , 25 which includes  $c(i,j)$  itself. By the properties of RSEFTF, each queue  $(i',j')$  can be scheduled at most  $nr_{i',j'} + 1$  times in any tau cell slots. Therefore, there can be at most  $S_i (3(n+m)/Sr_{i,j}) + n-1$  cells scheduled in the interval  $t-3(n+m)/S - (3(n+m)/S - 1)$ ,  $t-3(n+m)/S$  of  $3(n+m)/S$  cells slots to output  $j$  (not counting  $c(i,j)$ ). Similarly, for rates satisfying condition (1) there can be at most  $S_j (3(n+m)/Sr_{i,j}) + m-1$  cells at the input  $i$  which were scheduled in this interval. 30 Therefore, the total number of cells at time  $t-n$  which are either at input  $i$  or destined to output  $j$  which could still be at their inputs at the beginning of cell slot  $t-3(n+m-2)/S$  with time stamps less than or equal to  $t-3(n+m)/S$ , not counting  $c(i,j)$  is at most  $S_j (3(n+m)/Sr_{i,j}) +$

- 21 -

$S_i(3(n+m)/Sr_{i,j})+m+n-2 < 2 \times 3(n+m) \times S/3 + m + n - 2 = 3(m+n) - 2$ , falling short of the required number of  $3(m+n)-2$  needed to prevent  $c(i,j)$  from leaving by time  $t$ . Therefore, for any  $1 \leq S < 3$  there will not be enough cells with smaller or equal timestamps than  $c(i,j)$  to prevent  $c(i,j)$  from leaving its input before time  $t$ . QED.

CLAIMS

1. A method of providing bandwidth and delay guarantees in an input-buffered switch with a speed-up  $S$  having input channels and output channels for transferring cells there between, the method comprising:
  - 5 providing, to each of the input channels, per-output queues to buffer cells awaiting transfer to the output channels, each queue being associated with a respective input channel and an output channel, each queue having an assigned rate and an ideal service associated therewith;
  - 10 providing, to each of the input channels, a flow-level scheduler to provide bandwidth and delay guarantees, to schedule cells awaiting transfer at the input channels through flow queues
  - 15 providing, to each input channel, a rate controller to schedule, at a given cell slot, the per-output-channel queues in the input channel to which it is provided, the rate controller guaranteeing to each queue an amount of actual service that is within fixed bounds from the ideal service of the queue, the fixed bounds each being equal to a value of one;
  - 20 initializing state variables corresponding to the rate controller;
  - initializing a channel clock counter value to zero;
  - running each rate controller to select ones of the queues as scheduled at the given cell slot;
  - associating with indices corresponding to the selected ones of the queues timestamps equal to the current time; and
  - 25 for each matching phase where  $S$  is greater than or equal to two, performing arbitration processing to control transfer of the queued cells through the switch from the input channels to the output channels, the step of performing arbitration processing including:
    - performing a maximal match computation using the associated timestamps; and
    - indicating to each input channel the scheduled queues from which the input
    - 25 channel may transfer a cell;
    - advancing by one the channel clock counter value;
    - determining if the per-output queue rates have changed; and
    - returning to the step of running each rate controller.
  - 30 2. The method according to claim 1, wherein the maximal match computation comprises:
    - providing a set\_match set and a set\_queues set, the set\_match set being initialized to an empty set and the set\_queues set being initialized to the set of the associated timestamps;

- selecting a smallest of the associated timestamps stored in the set\_queues set;
- adding the selected associated timestamp to the set\_match set and removing the selected associated timestamp from the set\_queues set;
- deleting from the set\_queues set all remaining associated timestamps associated with per-output queues corresponding to either a same input channel or output channel as the selected associated timestamp;
- if the set\_queues set is empty, sending the indices of the queues corresponding to the timestamps in the set\_match set to the input channels to which they belong; and
- if the set\_queues set is not empty, then returning to the step of selecting.

10

3. The method according to claim 1, wherein running the rate controller comprises:
- for each per-output queue, maintaining the state variables to include a first and a second state variable, the first state variable corresponding to an ideal beginning time of the next cell of the per-output queue and the second state variable corresponding to an ideal finishing time of transmission of the next cell of the per-output queue;
- selecting as eligible all per-output queues having an ideal beginning time that is less than or equal to the current channel clock counter value;
- selecting as scheduled the eligible queue having a smallest finish time; and
- for the selected eligible queue, updating the first state variable with the ideal finish time and the second state variable with the ideal beginning time plus one divided by the assigned rate.

4. A method of providing bandwidth and delay guarantees in an input-buffered switch with a speed-up  $S$  having input channels and output channels for transferring cells therebetween, the method comprising:
- 25 providing, to each of the input channels, per-output queues to buffer cells awaiting transfer to the output channels, each queue being associated with a respective input channel and output channel, each queue having an assigned rate limited to 50% of a port bandwidth and an ideal service associated therewith;

- providing, to each of the input channels, a flow-level scheduler to provide bandwidth and delay guarantees and to schedule cells awaiting transfer at the input channels through flow queues and to assign the cells to ones of the per-output-channel queues;

- providing, to each input channel, a rate controller to schedule at a given cell slot the per-

output-channel queues in the input channel to which it is provided, the rate controller to guarantee to each queue an amount of actual service that is within fixed bounds from an ideal service of the queue, the fixed bounds each being equal to a value of one;

- initializing state variables corresponding to the rate controller;
- 5 initializing a channel clock counter value to zero;
- running each rate controller to select ones of the queues as scheduled at the given cell slot;
- associating with indices corresponding to the selected ones of the queues timestamps equal to the current time; and
- for each matching phase where  $S$  is greater than or equal to one but less than two,
- 10 performing arbitration processing to control transfer of the queued cells through the switch from the input channels to the output channels, the step of performing arbitration processing including the steps of:
  - performing a maximal match computation using the associated timestamps;
  - indicating to each input channel the scheduled queues from which the input
  - 15 channel may transfer a cell;
  - advancing by one the channel clock counter value;
  - determining if the per-output queue rates have changed; and
  - returning to the step of running each rate controller.
- 20 5. The method according to claim 4, wherein the maximal match computation comprises:
  - providing a set\_match set and a set\_queues set, the set\_match set being initialized to an empty set and the set\_queues set being initialized to the set of the associated timestamps;
  - selecting a smallest of the associated timestamps stored in the set\_queues set;
  - adding the selected associated timestamp to the set\_match set and removing the selecting
  - 25 associated timestamp from the set\_queues set;
  - deleting from the set\_queues set all remaining associated timestamps associated with per-output queues corresponding to either the same input channel or output channel as the selected associated timestamp;
  - if the set\_queues set is empty, sending the indices of the queues corresponding to the
  - 30 timestamps in the set\_match set to the input channels to which they belong; and
  - if the set\_queues set is not empty, then returning to the step of selecting.

6. The method according to claim 4, wherein running the rate controller comprises:
  - for each per-output queue, maintaining the state variables to include a first and a second state variable, the first state variable corresponding to an ideal beginning time of the next cell of the per-output queue and the second state variable corresponding to an ideal finishing time of transmission of the next cell of the per-output queue;
  - selecting as eligible all per-output queues having an ideal beginning time that is less than or equal to the current channel clock counter value;
  - selecting as scheduled the eligible queue having the smallest finish time; and
  - for the selected eligible queue, updating the first state variable with the ideal finish time and second state variable with the ideal beginning time plus one divided by the assigned rate.
7. An apparatus for providing bandwidth and delay guarantees in an input-buffered switch with a speed-up  $S$ , the input-buffered switch having input channels and output channels for transferring cells therebetween, the apparatus comprising:
  - per-output-channel queues in each of the input channels to buffer cells awaiting transfer to the output channels, each per-output-channel queue corresponding to a respective input channel and output channel, and having an assigned rate and an ideal service associated therewith;
  - a flow-level scheduler in each of the input channels to provide bandwidth and delay guarantees for scheduling cells awaiting transfer at the input channels through flow queues and to assign the cells to ones of the per-output-channel queues;
  - a rate controller, corresponding to each input channel, to schedule for a given cell slot, the per-output-channel queues in the input channel to which it corresponds, the rate controller to guarantee to each queue an amount of actual service that is within fixed bounds from an ideal service of the queue, the fixed bounds being equal to a value of one; and
  - an arbiter, responsive to the scheduling of queues by each rate controller, to control the processing of the queued cells through the switch from the input channels to the output channels at a speedup  $S$  equal to a number of phases per cell slot, where  $S$  is greater than or equal to two, the arbiter using a maximal match computation to choose ones of the scheduled queues from which a cell may be transmitted in each phase.
8. The apparatus according to claim 7, wherein the rate controller is located in the arbiter.

- 26 -

9. The apparatus according to claim 7, wherein the rate controller is loaded in the input channel to which it corresponds.
10. The apparatus according to claim 7, where S is greater than or equal to one but less than  
5 two with a load due to guaranteed flows limited to 50% of a port bandwidth.

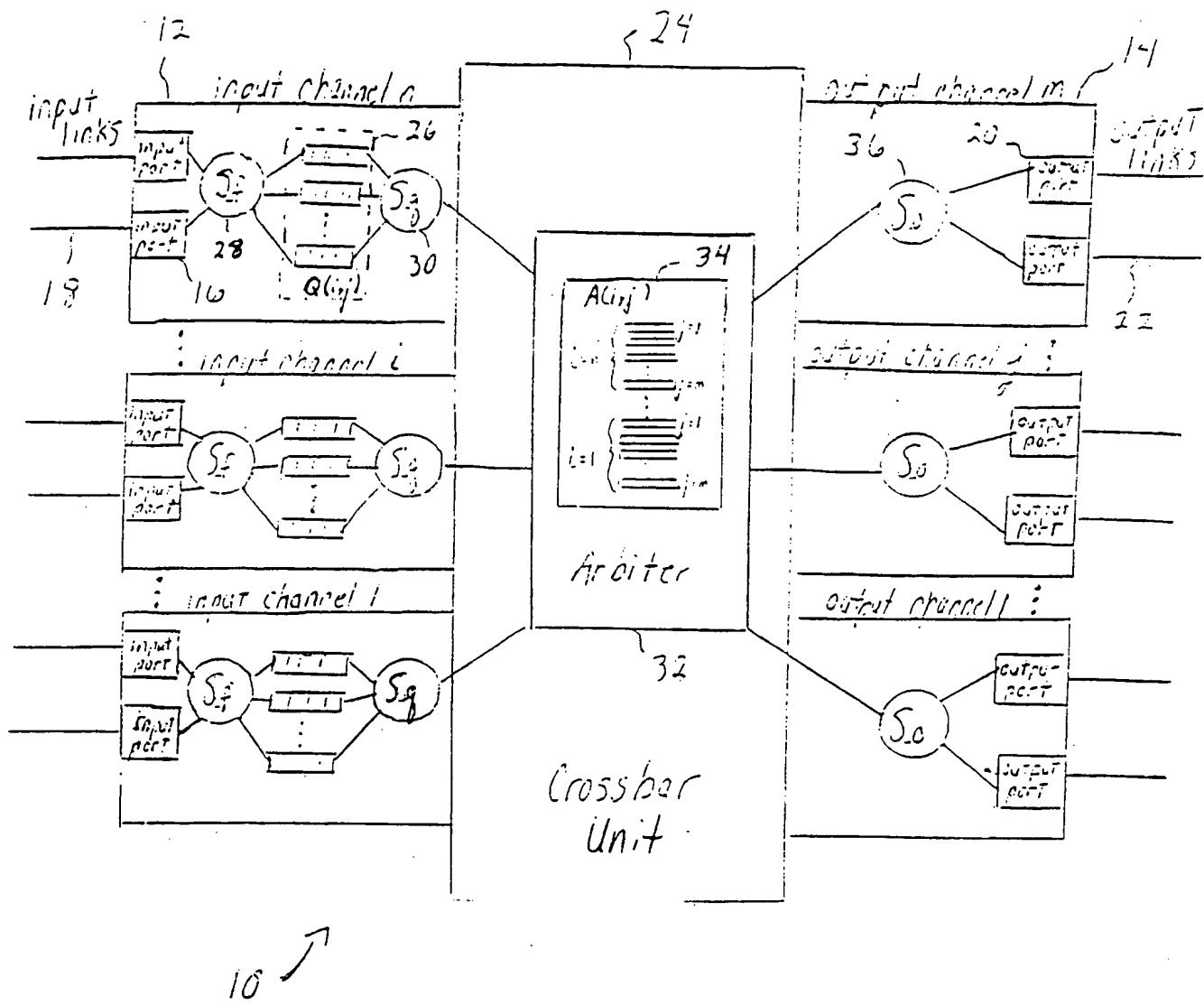


FIG. 1

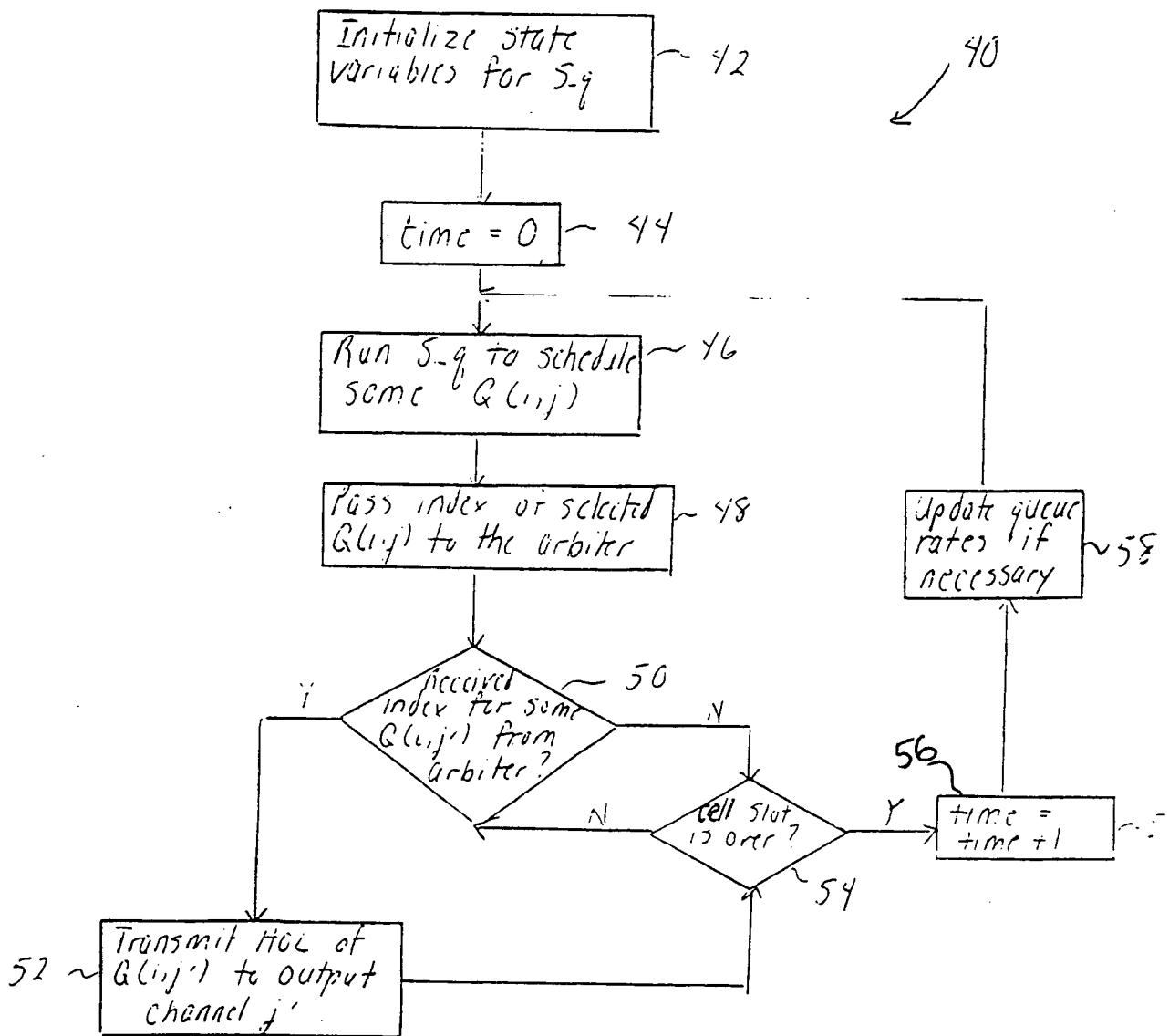


FIG. 2

3/4

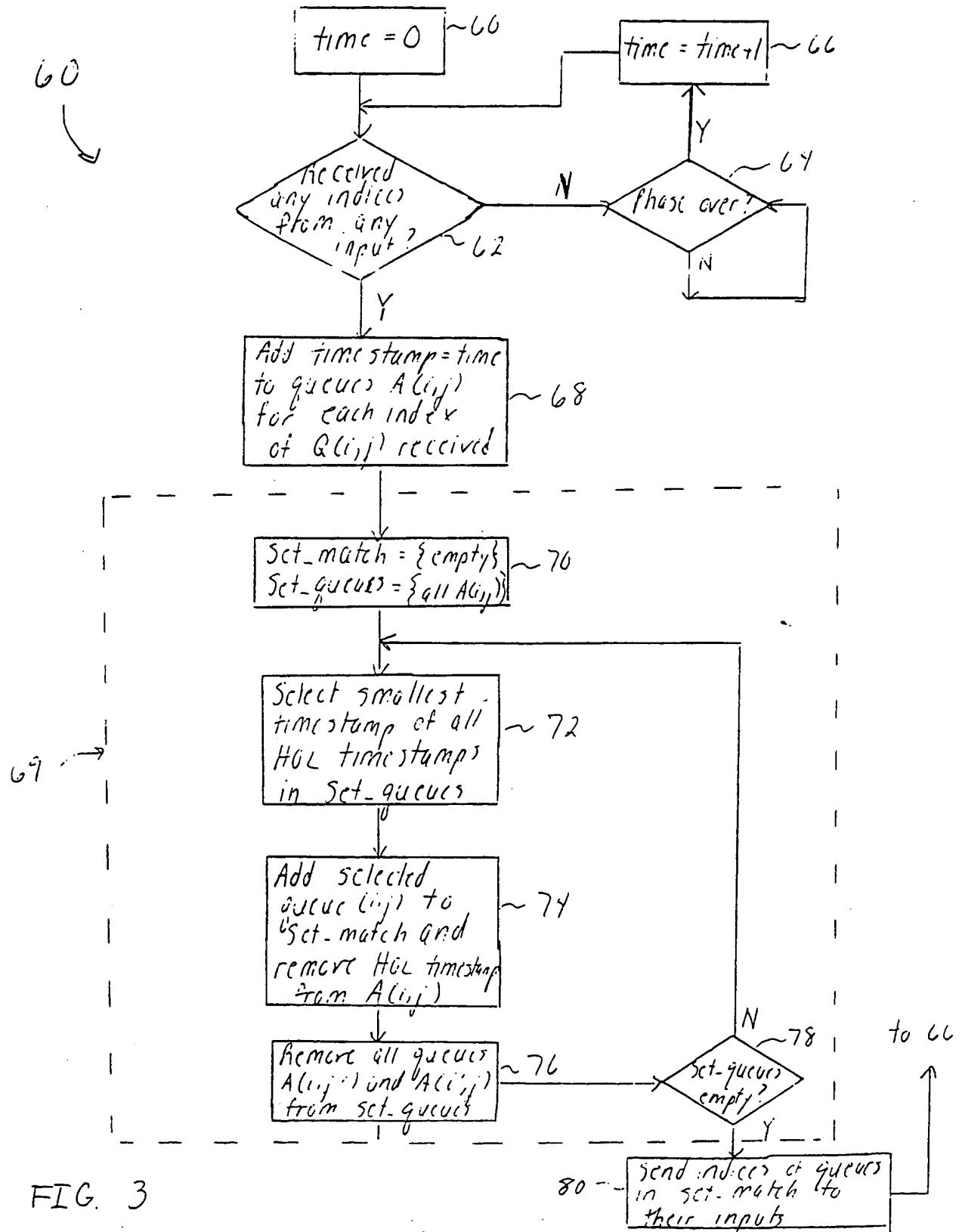


FIG. 3

4/4

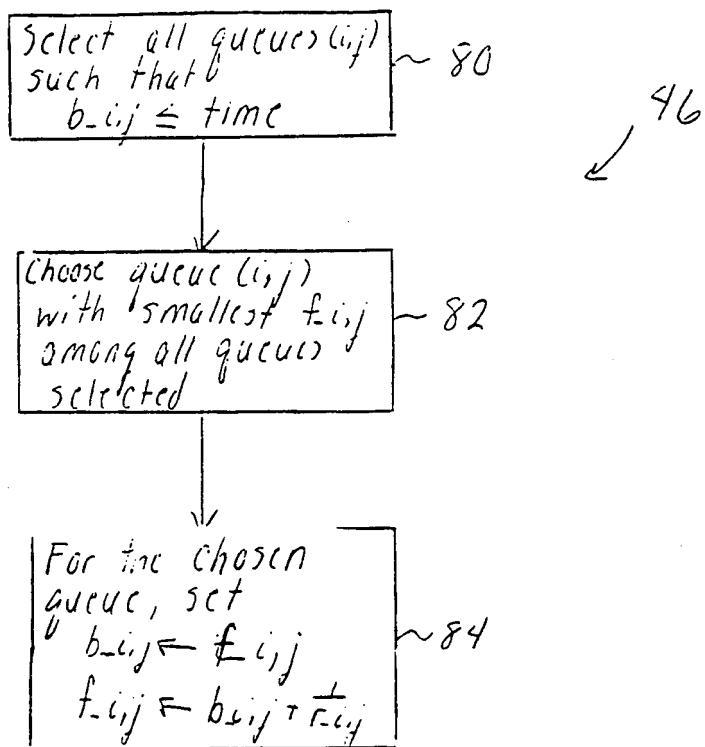


FIG. 4

## INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 99/00607

## A. CLASSIFICATION OF SUBJECT MATTER

IPC 6 H04Q11/04 H04L12/56

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 H04Q H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>LIU N H ET AL: "A NEW PACKET SCHEDULING ALGORITHM FOR INPUT-BUFFERED MULTICAST PACKET SWITCHES"            IEEE GLOBAL TELECOMMUNICATIONS CONFERENCE,            PHOENIX, ARIZONA, NOV. 3 - 8, 1997,            vol. 3, 3 November 1997, pages 1695-1699,            XP000737812            INSTITUTE OF ELECTRICAL AND ELECTRONICS            ENGINEERS            see paragraph I</p> <p>---</p> <p style="text-align: center;">-/--</p>	1, 4, 7



Further documents are listed in the continuation of box C.



Patent family members are listed in annex.

## ° Special categories of cited documents :

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

4 June 1999

Date of mailing of the international search report

15/06/1999

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
 NL - 2280 HV Rijswijk  
 Tel. (+31-70) 340-2040, Tx. 31 651 epo nl.  
 Fax: (+31-70) 340-3016

Authorized officer

Staessen, B

## INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US 99/00607

## C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MARK B L ET AL: "LARGE CAPACITY MULTICLASS ATM CORE SWITCH ARCHITECTURE" ISS '97. WORLD TELECOMMUNICATIONS CONGRESS. (INTERNATIONAL SWITCHIN SYMPOSIUM), GLOBAL NETWORK EVOLUTION: CONVERGENCE OR COLLISION? TORONTO, SEPT. 21 - 26, 1997, vol. 1, 21 September 1997, pages 417-423, XP000720547 ABE S ET AL see paragraph 2.1 ---	1,4,7
A	CHANG C -Y ET AL: "A BROADBAND PACKET SWITCH ARCHITECTURE WITH INPUT AND OUTPUT QUEUEING" PROCEEDINGS OF THE GLOBAL TELECOMMUNICATIONS CONFERENCE (GLOBECOM), SAN FRANCISCO, NOV. 28 - DEC. 2, 1994, vol. 1, 28 November 1994, pages 448-452, XP000488590 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS see abstract ---	1,4,7
A	PRABHAKAR B ET AL: "MULTICAST SCHEDULING FOR INPUT-QUEUED SWITCHES" IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, vol. 15, no. 5, 1 June 1997, pages 855-866, XP000657038 see paragraph I -----	1,4,7

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- BLACK BORDERS**
- IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- FADED TEXT OR DRAWING**
- BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- SKEWED/SLANTED IMAGES**
- COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- GRAY SCALE DOCUMENTS**
- LINES OR MARKS ON ORIGINAL DOCUMENT**
- REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**

THIS PAGE BLANK (USP10)